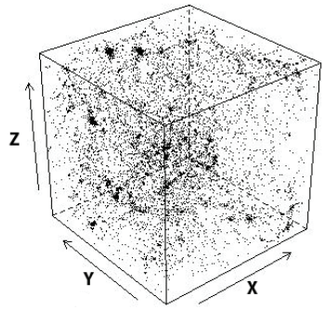
Detecting Distributional Differences in Simulated Particle Data Sets

Jeannette Figg, Joanne Wendelberger, Todd Graves, CCS-6; Jonathan Woodring, James Ahrens, CCS-7; Katrin Heitmann, ISR-1; Salman Habib, T-2

We explore methods to evaluate distributional differences in spatial point data sets. We used spatial point samples formed with two different sampling algorithms, as well as random samples from the same data set, as the basis for our comparison. Both visualization and quantitative techniques are described as a basis for comparing the spatial distributions of samples generated from the different sampling methods. Although the described quantitative methods show promise for future analysis, we were able to detect differences in our sampling algorithms using visualization methods alone. Visualizing the spatial density distribution along each of our axes allowed us to detect a distributional difference between the first sampling algorithm and the samples generated from the second sampling algorithm and the random samples. We hope to use spatial distributional functions and the properties of Voronoi and Delaunay tessellations to detect finer quantitative differences between the different samples in the future.

Fig. 1. A 3D map of the data point locations in one of our random samples.



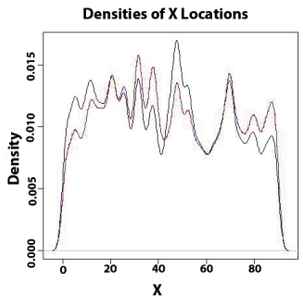


Fig. 2. The density along the x-axis for algorithm A (solid black line) compared to the density along the x-axis for algorithm B and the random samples (red dotted and blue solid lines). Notice the densities of algorithm B and the random samples appear almost identical.

Visualization and analysis methods are required to investigate massive datasets being collected or generated by simulations. For example, because cosmology particle data sets generated from both sky surveys and simulations have increased exponentially in size in the last decade, scalable methods to visualize and store these massive particle data sets are increasingly important [1]. Sampling the data to scale down the size of a data set presents a potentially scalable method to handle ever-increasing amounts of information. In addition to reducing the computational burden of a data set, sampling has the added advantage of retaining the original data points, whereas other methods often rely on compression or averaging. Here we describe some techniques we have investigated to evaluate different sampling algorithms for spatial point data.

We investigate two algorithms, referred to as algorithms A and B, which were written to be more computationally efficient than random sampling. Although based on the idea of random sampling, algorithms A and B introduce structure into the sampling process. Before samples are taken, algorithms A and B break the sample space into equal-density bins with each bin containing the same number of data points. The bins are then randomly sampled, with the same number of samples taken from each bin. As the desired sample resolution increases, increasing numbers of bins are used. Sampling algorithm A uses a bottom-up approach to decrease the sample resolution as it approaches the desired resolution, while sampling algorithm B uses a top-down approach to increase the resolution as it samples. Because the properties of random samples are relatively well known, we use a random sample of the data as a basis for evaluating the two algorithms on their statistical and visual properties. These comparisons formed our basis of evaluation for the sampling algorithms, as the algorithms were written to produce results similar to a pure random sample. We used thirty samples produced by each of the two algorithms, as well as thirty random samples for our primary analysis. Each sample contains 3D coordinate information for each data particle's location, as well as each particle's velocity in 3D. We assume that each particle has an identical mass. Using the techniques below, we detected substantial differences between samples formed with sampling algorithm A and the samples formed through a random sampling process.

Our two primary tools for investigating the samples formed from each algorithm were different visualizations of the samples and quantitative measurements of the samples' nonparametric properties. Because each sample contains in excess of 32 thousand points, basic 3D maps of the particle locations show no discernable difference between the random samples and samples formed with algorithms A and B. The maps do, however, display a clear indication of heavy clustering in some areas, prompting us to investigate the different density maps and representations of the data (Fig. 1). Although kernel density maps were not very helpful in distinguishing the samples, density curves along each visual and velocity axis were informative. For all samples produced with algorithms A and B, the density curves of the spatial locations of the points along the y- and z-axis, as well as the velocity density curves in

For more information contact Joanne Wendelberger at joanne@lanl.gov.

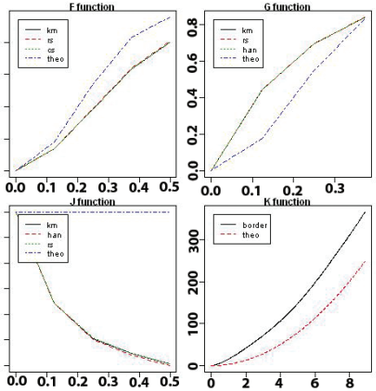


Fig. 3. The F, G, J, and K functions for one of our random samples. The blue dashed line represents the theoretical value for a Poisson process. All four functions show clear deviations from the Poisson process, correctly indicating that our sample contains clustering.

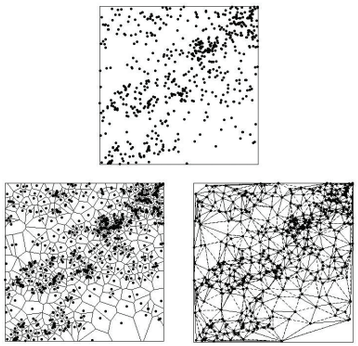


Fig. 4. Voronoi (bottom left) and Delaunay (bottom right) tessellations formed from a 2D section of one of the random samples (top).

all three dimensions appeared to lie right on top of the density curves from the random samples. The density curve along the spatial locations of points along the x-axis from sampling algorithm A, however, was visually distinct from the density curves generated from sampling algorithm B and the random samples (Fig. 2). Such a substantial deviation of sampling algorithm A from sampling algorithm B and the random samples seemed to indicate that algorithm A was dividing the x-axis in a way that distorted the density along that dimension.

In addition to the visualization techniques, we investigated several quantitative measures to aid our analysis of the different algorithms. Many of the quantitative measures used were compared to the Poisson point process, which displays random, independent scattering of points. Deviations from the Poisson characteristics can indicate either clustering or repulsion of points [2]. Because of our heavy clustering it is clear our samples differ from the Poisson process; however, we would like to investigate whether that difference changes between the algorithms and the random samples. In our future analysis, we would like to compare the distributions of several functions, the cumulative distribution of the empty space around a point (the F function), the cumulative distribution of the distance *r* of the nearest neighbor to a point (the G function), the ratio of the F and G functions [3], and the probability of finding two points within a distance *r* of each other (g) [4], from each of our sampling algorithms. The distributions of the four functions generated from one of our random samples are shown in Fig. 3 for the random sample. As expected, the functional characteristics from all three sample algorithms differed significantly from the Poisson process. We hope the comparison of the functions generated from the three algorithms will provide further insight into the quantitative similarities or differences between the sampling methods.

In addition to the techniques used above, several techniques utilizing the properties of the Delaunay and Voronoi tessellations seem promising for future analysis. Although often considered a visualization tool, the Voronoi and Delaunay tessellations can also provide quantitative information to our sample analysis. The two tessellation techniques are each others' analogue, where the Voronoi tessellation divides the sample space into cells each containing one data point as its nucleus and the Delaunay tessellation connects the nuclei of these cells (Fig. 4). The geometrical properties of the Voronoi tessellation, such as the distribution of the number of vertices per cell, the volume of the cell and the surface area of the cell can be measured quantitatively. These quantitative measures have the potential to provide information on the clustering and voids of the point pattern from which the Voronoi tessellation was formed [5]. With the Delaunay tessellation, a Delaunay Tessellation Field Estimator (DTFE) can be constructed as a means to transform the discrete points contained in the sample into a continuous field [6]. Because the Delaunay tessellation is so sensitive to local point densities, the DTFE can be used as a local density estimate [7]. The density estimate could then be used to further characterize the clusters and voids present in each sample.

Through early analysis and visualization, we were able to detect a difference in the way algorithm A distributes sampling along the x-axis. We hope that with continued analysis using our proposed quantitative techniques we will be able to detect finer differences in the distributions of samples from simulations.

[1] Borne, K., "Astroinformatics: A 21st Century Approach to Astronomy," *Astro2010 Position Papers*, **6** (2010).
[2] Illian, J., et al., *Statistical Analysis and Modelling of Spatial Point Patterns,* John Wiley & Sons, England (2008).
[3] van Lieshout, M.N.M. and A.J. Baddeley, *Statistica Neerlandica* **50,** 344 (1996).
[4] Ripley, B.D., *J Roy Stat Soc B Stat Meth* **39,** 172 (1977).
[5] van de Weygaert, R., *Astron Astrophys* **283,** 361 (1994).
[6] Schaap, W. and R. van de Weygaert, "The Delaunay Tessellation Field Estimator," http://dissertations.ub.rug.nl/FILES/faculties/science/2007/w.e.schaap/c2.pdf (2007).
[7] van de Weygaert, R. and W. Schaap, *Physics* **665**, 291 (2009).

Funding Acknowledgments
LANL Laboratory Directed Research and Development; Advanced Simulation and Computing (ASC)